Ch-8 Advance Topics:

Clustering:

→ Cluster is a group of objects that belongs to the same class.

→ Clustering: D.M technique used to place elements into related groups without advanced knowledge of group definition.

→ applications: - Marketing.
- Libraries (similar books).
- Insurance.
- City planning
- Earthquake studies
- www.

2
SUNDAY

→ requirements: → Scalability
→ Ability to deal with different kinds of attributes
→ Ability to deal with noisy data.
→ Interpretability.

NOTES

JULY						2017
S	M	T	W	T	F	S
						1
30	31					
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

* Spatial Mining:

→ Spatial data mining is based on geograp
analysis

→ Analysts use geographical or spatial info
to produce intelligence

→ Specific techniques are required to convert these
data into useful format.

→ Task: Search for 'Spatial Patterns'

* Web Mining:

→ automatically extract information from web pages /
documents & services.

→ information discovered : web activity, server logs
and browser activity tracking

→ Types: - Web Content Mining : within Page info (Clustering
- Web Structure Mining : inter Connections of Page (links
- Web Usage Mining : patterns and trends
to improve structure.

eg. logs, app ~~serve~~ logs

JULY 2017

S	M	T	W	T	F	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

NOTES

* Text Mining :

→ text data mining is process of deriving high quality information from text.

→ Areas :

→ Information Retrieval :

→ ability to query a comp. system to return relevant results.

→ Natural Language Processing :

→ NLP is one of the challanging problems.

→ Study of human language so it understands natural language.

* **Big Data:** → data which is large in quantity

Volume :
Velocity :
Veracity :
Verity :

* **Hadoop:** distributed processing framework that manages data processing and storage for big data app. running in clustered system.

→ Provides Massive Storage.

→ **Nodes:**

→ 1) Name Node
2) Secondary name node } master
3) Job Tracker

→ 4) Data Node } slave
5) Task Tracker

*Hadoop - Distributed File System (HDFS)

-> stores data across multiple machines without prior organization.

* Map Reduce:

-> parallel processing software framework.

-> Steps: i) Mapper: Partitioning.
ii) Reducer: Shuffle, sort, combine and produce o/p.

* YARN (Yet Another Resource Navigator).

-> provides resource management for processes running on hadoop.

=> Master node for data storage => Name node.

=> Master node for parallel processing (Map Reduce). => Job Tracker.

* Name node:

→ all files and directories in HDFS namespace are represented on the namenode by Inodes.

→ Contains attributes like permissions, modification timestamp, disk space quota, namespace quota, access time.

→ maps entire file system structure into memory.

→ fs image: inodes and the list of block which define the metadata.

→ edits: modifications performed on the content of fsimage.

* Data Node:

→ Manages state of HDFS node and interacts with blocks.

→ Data node can perform CPU insentive jobs. ml tasks, I/O jobs, data import export, indexing etc.

→ On startup every data node connects to the name node & performs handshake for verificatio

→ Datanode sends heartbeat to namenode every 3 seconds to confirm that the data node is operating and the block replicas it hosts are available.

→ Verifies block replicas.

→

→ Why hadoop?

→ Scalability

→ Flexibility

→ Computing power

→ fault tolerance (protection again hardware failures.)

→ low cost

→ Storage and processing speed.

Challangs:-

- Map Reduce not always suitable. (not for iterative tasks.)
- Data Security: (to make it secure kerberos authentication protocol is used).
- Difficult for new programmers.
- lacking of tools for data quality and Standardiza